Detecting Adversarial Faces Using Only Real Face Self-Perturbations

Qian Wang[1], **Yongqin Xian**[2], **Hefei Ling**[1,*], **Jinyuan Zhang**[3], **Xiaorui Lin**[3], **Ping Li**[1], **Jiazhong Chen**[1], **Ning Yu**[4]
[1]Huazhong University of Science and Technology, Wuhan, China
[2]Google, Switzerland
[3]Software Development Center, Industrial and Commercial Bank of China
[4]Salesforce Research, USA
[1]{yqwq1996, lhefei, lpshome, jzchen}@hust.edu.cn, [2]yxian@google.com,
[3]{zhangjy, linxr}@sdc.icbc.com.cn, [4]ning.yu@salesforce.com

Abstract

Adversarial attacks aim to disturb the functionality of a target system by adding specific noise to the input samples, bringing potential threats to security and robustness when applied to facial recognition systems. Although existing defense techniques achieve high accuracy in detecting some specific adversarial faces (adv-faces), new attack methods especially GAN-based attacks with completely different noise patterns circumvent them and reach a higher attack success rate. Even worse, existing techniques require attack data before implementing the defense, making it impractical to defend newly emerging attacks that are unseen to defenders. In this paper, we investigate the intrinsic generality of adv-faces and propose to generate pseudo adv-faces by perturbing real faces with three heuristically designed noise patterns. We are the first to train an adv-face detector using only real faces and their self-perturbations, agnostic to victim facial recognition systems, and agnostic to unseen attacks. By regarding adv-faces as out-of-distribution data, we then naturally introduce a novel cascaded system for adv-face detection, which consists of training data self-perturbations, decision boundary regularization, and a max-pooling-based binary classifier focusing on abnormal local color aberrations. Experiments conducted on LFW and CelebA-HQ datasets with eight gradient-based and two GAN-based attacks validate that our method generalizes to a variety of unseen adversarial attacks.

1 Introduction

[1]Deep neural networks have been widely used in many tasks [Ou *et al.*, 2017; Shi *et al.*, 2020; Zhang *et al.*, 2021], and have achieved remarkable success in facial recognition systems (FRS) [Deng *et al.*, 2019] with a wide range of real-world applications such as online payment [Avarikioti *et al.*, 2019] and financial management [Dumbre *et al.*, 2016]. However, deep neural networks are known to be vulnerable to adversarial attacks [Goodfellow *et al.*, 2014; Kurakin *et al.*, 2016; Madry *et al.*, 2017], making commercial FRSs unreliable.

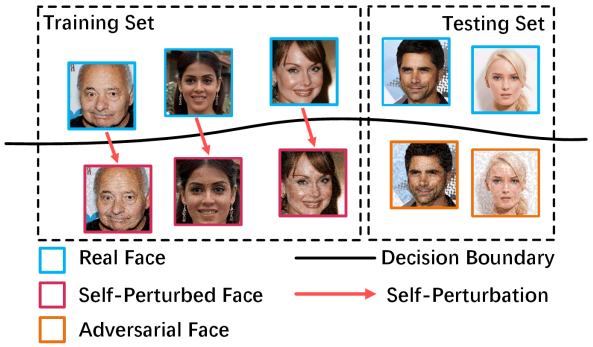

Figure 1: Trained only on real faces and their self-perturbed faces, the detector learns a generic representation of adversarial faces produced by unseen attacks.

Some research attempts to defend against adversarial attacks by adversarial example detection techniques [Feinman *et al.*, 2017; Ma *et al.*, 2018; Tian *et al.*, 2021] which filter out adversarial samples before feeding them into protected systems. However, they tend to overfit to known attacks and do not generalize well to unseen advanced attacks [Wu *et al.*, 2021; Yang *et al.*, 2021]. In specific, GAN-based adversarial attacks such as AdvMakeup [Yin *et al.*, 2021] and AMT-GAN [Hu *et al.*, 2022] are recently developed to generate more natural adversarial faces (adv-faces) by adding completely different noise patterns to face images in contrast to traditional gradient-based attacks such as FGSM [Goodfellow *et al.*, 2014], and are able to circumvent detection. Therefore, a plug-and-play method of detecting adv-faces with improved generalization performance over both unseen gradient-based attacks and unseen GAN-based attacks is highly demanded.

Due to the insight that adversarial samples are all modified from real samples [Ma *et al.*, 2018], one direct way to encourage models to learn generic representations for detecting adv-faces is to train models with synthetic data [Shiohara and Yamasaki, 2022], forming a decision boundary wrapping

*Corresponding author
[1]Code at https://github.com/cc13qq/SAPD

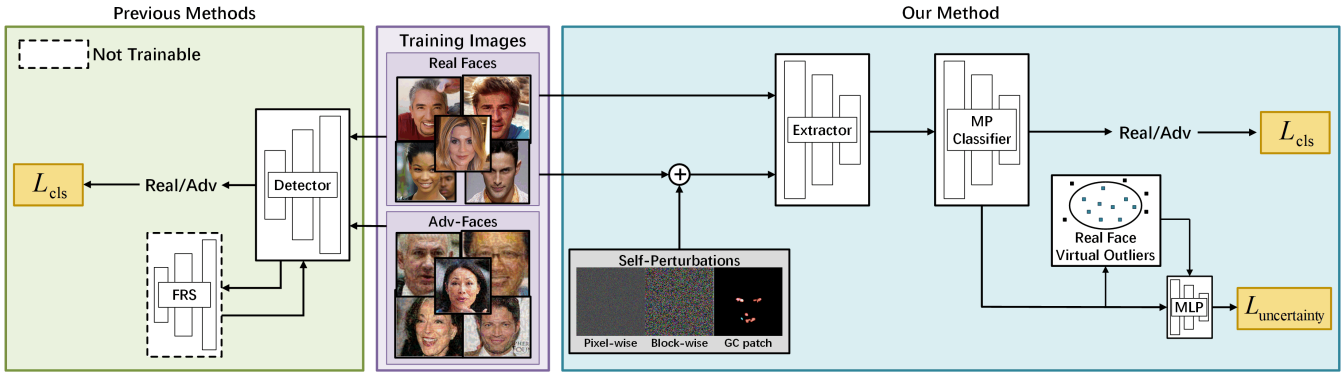

Figure 2: Overview of training pipeline. We design three kinds of self-perturbations to generate synthesized pseudo-adv faces. "GC patch" denotes a gradient color patch for GAN-based attack. "MP Classifier" denotes Max-Pooling Classifier which focuses on abnormal local color aberrations. By synthesizing virtual outliers of real faces, we incorporate an uncertainty loss to regularize the decision boundary to enhance detection performance on adv-faces. Previous methods require pre-computed adv-faces and access permission to the victim FRS for training. By contrast, our method is agnostic to both attacks and FRS. Using only real faces with self-perturbations in the training phase, our method is able to detect all kinds of adv-faces from unseen attacks without modifying or visiting FRS.

the synthetic subspace. To achieve this, we investigate the intrinsic generality of adv-faces generated by gradient-based attacks and GAN-based attacks respectively, and propose two assumptions: (1) Gradient-based attacks have a resemblance in noise pattern on account of the consistency of the basic attack algorithm which uses gradients to construct adversarial examples. (2) GAN-based attacks aim to simultaneously change the prediction of FRSs while preserving visual quality. In the meantime, FRSs pay the most attention to high-frequency regions [Li *et al.*, 2020], which induces GANs to modify these regions. Therefore, GAN-made adv-faces have manipulated clues like abnormal color aberrations in high-frequency regions. If universal noise patterns that cover all kinds of adversarial noises are attainable, a set of pseudo adv-faces could be produced, and we can train a model with them to learn generic representations for adv-faces.

Based on the above assumptions, we propose three kinds of real face self-perturbations synthesized with pseudo noise patterns which summarize adversarial noises of gradient-based attacks and GAN-based attacks, and a max-pooling-based classifier focusing on capturing abnormal local color aberration. Unlike previous methods, this data-augmentation-like method is proposed from a novel and more intrinsic perspective by investigating the generality of adversarial noise patterns, to detect all recent adv-faces from unseen attacks, protecting FRSs without access to them. Trained on only real faces and their self-perturbations, as shown in Figure 1, our model is able to detect all the recent adv-faces. This new framework simultaneously ensures detection performance on unseen adversarial attacks and portability in use, without any access to protected systems.

Although we are trying to make self-perturbed faces as general as possible, the trained network still overfits to in-domain data and might fail to classify adv-faces far away from the distribution of self-perturbed faces. Regarding adv-faces as out-of-distribution (OOD) data from a shifted data manifold [Deng *et al.*, 2021], we incorporate a regularization term to narrow the model's decision boundary of real face class during training, and naturally introduce a novel cascaded system for adv-face detection, which consists of training data self-perturbations, decision boundary regularization, and max-pooling-based binary classifier. We evaluate our approach on LFW and CelebA-HQ datasets with eight gradient-based attack methods and two GAN-based attack models. Experiment results validate our assumptions and demonstrate the reliability of our detection performance on adv-faces from unseen adversarial attack methods.

Our contributions are summarized in four thrusts:

- By investigating the intrinsic similarities among varying adv-faces, we propose two assumptions that (1) different adversarial noises have resemblances, and (2) color aberrations exist in high-frequency regions. We empirically validate the assumptions which in turn indicate universal noise patterns are attainable for all the recent adv-faces.
- Based on our assumptions, we propose three kinds of real face self-perturbations for gradient-based adv-faces and GAN-based adv-faces. We are the first to train an adv-face detector using only real faces and their self-perturbations, agnostic to victim FRSs and agnostic to adv-faces. This enables us to learn generalizable representations of adv-faces without overfitting to any specific one.
- We then naturally introduce a novel cascaded system for adv-face detection, which consists of training data self-perturbations, decision boundary regularization, and a max-pooling-based binary classifier focusing on abnormal local color aberrations.
- Evaluations are conducted on LFW and CelebA-HQ datasets with eight gradient-based attacks and two GAN-based attacks, demonstrating the consistently improved performance of our system on adv-face detection.

2 Related Work

Adversarial attacks Adversarial attacks [Goodfellow *et al.*, 2014] aim to disturb a target system by adding subtle noise to input samples, while maintaining imperceptibility from human eyes. In contrast to previous works

[Madry *et al.*, 2017; Kurakin *et al.*, 2016], DIM [Wu *et al.*, 2021] and TIM [Dong *et al.*, 2019] are improved attack algorithms with enhanced black-box attack accuracy and breaching several defense techniques. Focusing on breaking through FRSs, TIPIM [Yang *et al.*, 2021] generates adversarial masks for faces to conduct targeting attacks while remaining visually identical to the original version for human beings. Recently, GAN-based attack models [Yin *et al.*, 2021; Hu *et al.*, 2022] are presented to generate adversarial patches and imperceptible adv-faces, bringing new challenges to adv-faces detection. In this paper, we proposed a simple yet effective detection method against all the above attacks while being blind to them during training.

Adversarial example detection One of the technical solutions for protecting DNNs from adversarial attacks is adversarial example detection [Feinman *et al.*, 2017], aiming to filter out adversarial inputs before the protected system functions. Remedying the limitations of the previous method, LID [Ma *et al.*, 2018] is proposed for evaluating the proximity of an input to the manifold of normal examples. Incorporated with wavelet transform, SID [Tian *et al.*, 2021] is able to transform the decision boundary and can be collaboratively used with other classifiers. While all of these methods focus on utilizing features subtracted by FRS backbone and show effective performance in detecting adversarial examples, they require modifying or visiting the protected system and do not generalize well to new-type attacks like GAN-based AMT-GAN [Hu *et al.*, 2022] and AdvMakeup [Yin *et al.*, 2021]. To address the problems above, our method is proposed to detect all recent adv-faces from unseen attacks, protecting FRSs without access to them.

Out-of-distribution (OOD) detection OOD detection techniques [Girish *et al.*, 2021] have been widely used in image classification tasks [Yu *et al.*, 2019], trying to recognize examples from an unknown distribution. ODIN [Liang *et al.*, 2017] uses temperature scaling and tiny perturbations to the inputs to separate the in-distribution (ID) and OOD images. Lee et al. [Lee *et al.*, 2018] uses the Mahalanobis distance to evaluate the dissimilarity between ID and OOD samples. By sampling and synthesizing virtual outliers from the low-likelihood regions, VOS [Du *et al.*, 2022] adaptively regularizes the decision boundary during training. Sun et al. [Sun *et al.*, 2021] introduce a simple and effective post hoc OOD detection approach utilizing activation truncation. Regarding adv-faces as OOD data, we leverage an uncertainty regularization term to narrow the decision boundary of real face class during the training phase to boost the accuracy of adv-faces detection.

3 Methodology

We propose a plug-and-play cascaded system for adv-faces detection method which consists of training data self-perturbations, decision boundary regularization, and a max-pooling-based binary classifier focusing on abnormal local color aberrations, agnostic to unseen adversarial attacks, and agnostic to victim FRSs. Training pipeline as shown in Figure 2. We propose to synthesize diverse self-perturbed faces by adding three noise patterns to real face images, summarizing adv-faces generated from gradient-based and GAN-based attacks. A convolutional neural network is learned via real faces and self-perturbed faces, regularized by uncertainty loss, to distinguish real and adversarial faces. As no attack method is observed during training, the resulting network is not biased to any attack, yielding generic and discriminative representations for detecting unseen adv-faces. Therefore, in the testing phase, the embeddings are sent to the learned Max-Pooling Classifier to accomplish prediction.

Algorithm 1 Self-perturbation for gradient-based attack

Input: A empty perturbation matrix $\boldsymbol{\eta}^p \in \mathbb{R}^{H\times W\times 3}$ with the same shape of real face image $\mathbf{x}^r$.
Parameter: Max perturbation magnitude ϵ, pattern mode.
Output: Self-perturbed face image $\mathbf{x}^p$.

A random direction matrix $R = \{\vec{r}_{ij}\} \in \mathbb{R}^{H\times W\times 3}$.
for $\vec{\eta}_{ij}$ **in** $\boldsymbol{\eta}^p$ **do**
  Select random noise value α.
  if pattern mode **is** 'point-wise' **then**
    $\vec{\eta}_{ij} := \alpha \cdot \vec{r}_{ij}$.
  else if pattern mode **is** 'block-wise' **then**
    Select a random neighborhood A_{ij} of $\vec{\eta}_{ij}$.
    for $\vec{\eta}_{ijk}$ **in** A_{ij} **do**
      $\vec{\eta}_{ijk} := \alpha \cdot \vec{r}_{ij}$.
    end for
  end if
end for
Clip perturbation $\boldsymbol{\eta}^p$ using Equation 2.
Generate self-perturbed face $\mathbf{x}^p$ using Equation 3.
return $\mathbf{x}^p$

3.1 Real Face Self-Perturbations

Self-perturbation for gradient-based attack

Given a real face image $\mathbf{x}^r$, adversarial attack generates an adversarial image $\mathbf{x}^a$ by adding a perturbation image $\boldsymbol{\eta}$ to $\mathbf{x}^r$ [Goodfellow *et al.*, 2014]:

$$\mathbf{x}^a = \mathbf{x}^r + \boldsymbol{\eta}, \tag{1}$$

where $||\boldsymbol{\eta}|| \leq \epsilon$, and ϵ called perturbation magnitude.

We observe that a binary classifier trained on real faces and adv-faces generated by FGSM is able to classify a part of attack images generated from other attack methods as shown in Table 2. This generalization in classification means that attack noises have intrinsic similarities. And it means once we master all the similarities we master the attacks, even for unseen attacks. This leads to non-trivial designs for our self-perturbation image $\boldsymbol{\eta}^p$.

The noise value of neighbor points in $\boldsymbol{\eta}^p$ may be the same or different. From this perspective, we introduce point-wise and block-wise noise patterns for gradient-based attacks. As presented in Algorithm 1, we perturb each point in the point-wise pattern and each block in the block-wise pattern in a stochastic direction, where blocks are random neighborhoods of a set of scattered points. The generated perturbation image $\boldsymbol{\eta}^p$ is constrained in l^∞ norm, and clipped according to ϵ,

$$\boldsymbol{\eta}^p = \mathrm{Clip}_{[-\epsilon,\epsilon]}(\boldsymbol{\eta}^p). \tag{2}$$

Algorithm 2 Self-perturbation for GAN-based attack

Input: Real face image $\mathbf{x}^r$. A empty perturbation matrix $\boldsymbol{\eta}^p = \{\eta_{ij}\}$ with the same shape of $\mathbf{x}^r$.
Parameter: Max perturbation magnitude ϵ, face landmarks, high-frequency threshold γ, and empty high-frequency pixel set H.
Output: Self-perturbed face image $\mathbf{x}^p$.
Function: $\mathrm{value}(x)$ measures pixel value of x, $\mathrm{neighborhood}(x)$ is a random neighborhood of x.

```
Obtain gradient image of x^r using Sobel operator.
Obtain high-frequency convex hull from gradient image according to landmarks.
for x_ij in convex hull do
    if value(x_ij) < γ then
        η_ij join to H.
    end if
end for
Random select a subset H_s of H.
for η_ij^H in H_s do
    Generate a random gradient color patch P.
    neighborhood(η_ij^H) := P.
end for
Clip perturbation η^p using Equation 2.
Generate self-perturbed face x^p using Equation 3.
return x^p
```

And self-perturbed face is calculated as

$$\mathbf{x}^p = \mathbf{x}^r + \boldsymbol{\eta}^p. \tag{3}$$

Self-perturbation for GAN-based attack

Spectrums of real and fake images distribute differently in high frequency [Luo *et al.*, 2021], such as eyes, nose, and mouth. This distribution difference should become more intense because GAN-made adv-faces aim to change the classification results of FRSs meanwhile maintain perceptual invariance. This leads to color aberration and boundary abnormality in high-frequency regions on account of that FRSs pay the most attention to these regions. Focusing on producing natural and imperceptible attack noise, these abnormal color aberrations should have gradient or blurred boundaries. In such a perspective, we use gradient color patches to act as self-perturbation.

As presented in Algorithm 2, we first obtain the gradient image of a real image using Sobel operator [Duda and Hart, 1973] and convex hull of the high-frequency area according to facial landmarks. The pixels in the convex hull are selected as high-frequency pixels if their values surpass a threshold γ. Then we generate a few gradient color patches through a series of affine transformations. Finally, the pseudo adv-faces are generated by adding gradient color patches to a random set of high-frequency pixels.

After crafting a part of real faces to self-perturbed faces as negative samples, we label other real faces as positive samples and train a well-designed backbone network such as XceptionNet [Chollet, 2017] in a binary classification manner.

3.2 Max-Pooling Classifier

In view of abnormal speckles in GAN-based adv-faces are always tiny and are not easy to be observed, we proposed Max-Pooling Classifier (MPC) to capture abnormal local color aberrations. MPC produces classification scores and predicts whether input images are real or adversarial.

A typical backbone network includes a feature extractor $f(\cdot;\boldsymbol{\varphi})$ composed of several convolutional blocks, and a classifier $g(\cdot;\boldsymbol{\phi})$ comprising an average pooling layer, an activation layer, and a fully connected layer [Chollet, 2017]. To detect adv-faces with abnormal local color aberrations, a simple derivation is to divide an image into several rectangle areas, and if abnormal color aberration occurs in any of the rectangle areas we classify this image as adv-face.

Input an image $\mathbf{x}$ to feature extractor $f(\cdot;\boldsymbol{\varphi})$, it produces a feature map $M(\mathbf{x}) \in \mathbb{R}^{N\times N\times d}$ composed of $N \times N$ features, where d is the embedding size, and each feature is corresponding to a rectangle area i.e. the receptive field. Features are sent to the activation function and fully connected layer respectively and produce $N \times N$ logits. We take the max-pooling of these logits as the final logit and obtain prediction score $S_{\text{cls}} \in \mathbb{R}^2$ by softmax operation. Classification of input image $\mathbf{x}$ is computed by argmax function:

$$G_{\text{cls}}(\mathbf{x}) = \mathrm{argmax}(S_{\text{cls}}), \tag{4}$$

where $G_{\text{cls}}(\mathbf{x}) = 1$ indicates real face and $G_{\text{cls}}(\mathbf{x}) = 0$ indicates adv-face.

3.3 Decision Boundary Regularization

Although we are trying to make self-perturbed faces as general as possible, the trained network still overfits to ID data and might fail to classify faces far away from the distribution of negative samples. Regarding adv-faces as OOD data, we incorporate a regularization term [Du *et al.*, 2022] to our work to enhance adv-face detection performance.

Assuming the feature representation of real faces forms a multivariate Gaussian distribution, we sample virtual outliers $\mathcal{V} \subset \mathbb{R}^m$ from the ε-likelihood region of the estimated class-conditional distribution:

$$\mathcal{V} = \left\{ \mathbf{v} \mid \frac{\exp\left(-\frac{1}{2}\left(\mathbf{v}-\widehat{\boldsymbol{\mu}}\right)^\top \widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{v}-\widehat{\boldsymbol{\mu}}\right)\right)}{(2\pi)^{m/2}|\widehat{\boldsymbol{\Sigma}}|^{1/2}} < \varepsilon \right\}, \tag{5}$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the estimated mean and covariance using latent features of real faces.

The uncertainty loss regularizes the model to produce a low OOD score for ID data and a high OOD score for the synthesized outliers, and narrows the decision boundary of real face class to boost the performance of adv-face detection:

$$\begin{aligned}\mathcal{L}_{\text{uncertainty}} = {} & \mathbb{E}_{\mathbf{v}\sim\mathcal{V}}\left[\log(\exp^{-\phi(-f(\mathbf{v};\theta))}+1)\right] \\ & + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}^r}\left[\log(\frac{1}{\exp^{-\phi(-f(\mathbf{x};\theta))}}+1)\right],\end{aligned} \tag{6}$$

where $\mathcal{D}^r$ represents distribution of real face, $f(\cdot;\theta)$ is a linear transformation function, and $\phi(\cdot)$ is a nonlinear MLP function. The learning process shapes the uncertainty surface, which predicts a high probability for ID data and a low probability for virtual outliers $\mathbf{v}$.

	Study	Method	Detect Attacks	Attack-Agnostic	FRS-Agnostic
AD	LID [Ma *et al.*, 2018]	Local intrinsic dimensionality	Gradient-based	×	×
AD	SID [Tian *et al.*, 2021]	Wavelet transform	Gradient-based	×	×
FD	Luo et al. [Luo *et al.*, 2021]	SRM convolution + DCMA	GAN-based	×	✓
FD	He et al. [He *et al.*, 2021]	Re-Synthesis Residuals	GAN-based	×	✓
OOD	ODIN [Liang *et al.*, 2017]	Softmax score	Gradient-based + GAN-based	✓	×
OOD	MD [Lee *et al.*, 2018]	Mahalanobis distance.	Gradient-based + GAN-based	✓	×
OOD	ReAct [Sun *et al.*, 2021]	Rectified truncation	Gradient-based + GAN-based	✓	×
	Ours	Self-Perturbation	Gradient-based + GAN-based	✓	✓

Table 1: Baselines used in our study. "AD" and "FD" denotes adversarial example detection and face forgery detection respectively. "OOD" denotes OOD detection. "Attack-Agnostic" means that attacks are unseen and no adv-faces are required for training. "FRS-Agnostic" means that victim FRS are unknown and access is forbidden. Only our method does not rely on pre-computed adv-faces or victim FRS.

The training objective combines the real-adv classification loss and the regularization term:

$$\min \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}(\mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{uncertainty}}), \tag{7}$$

where $\mathcal{D}$ represents the distribution of training data. β is the weight of the regularization $\mathcal{L}_{\text{uncertainty}}$, and $\mathcal{L}_{\text{cls}}$ is the Cross-Entropy classification loss [Zhang and Sabuncu, 2018].

4 Experiment

To validate the detection performance of our approach on adv-faces generated by various adversarial attack methods, we conduct extensive empirical studies on two datasets in this section. We validate our assumptions by investigating the intrinsic similarity of adv-faces and compare our method to baselines from three research streams. After that, we analyze our approach through a series of ablation studies.

General setup During model training, real faces are used as positive samples, and only self-perturbations of real faces are used as negative samples. Any adv-faces are agnostic. During testing, all the negative samples are adv-faces. In detail, half of the real faces in the training phase are labeled 1 as the positive samples, and others are self-perturbed and labeled 0 as the negative samples. In the testing phase, all of adv-faces are labeled 0. For evaluating the performance of detectors, we choose the widely used AUC score as the main metric.

Datasets Face images in this work are sampled from LFW [Gary *et al.*, 2007] and CelebA-HQ [Karras *et al.*, 2017] datasets. LFW contains 13,233 face images of 5,749 subjects. Subjects with at least two face images take part in adv-face producing. The first image of each subject is regarded as a reference and the others are sampled to produce adv-faces which includes 7,484 images. CelebA-HQ is a high-resolution subset of CelebA, containing 30,000 images.

Attack methods We employ eight recent gradient-based adversarial attack methods FGSM [Goodfellow *et al.*, 2014], BIM [Kurakin *et al.*, 2016], PGD [Madry *et al.*, 2017], RFGSM [Tramèr *et al.*, 2018], MIM [Dong *et al.*, 2018], DIM [Wu *et al.*, 2021], TIM [Dong *et al.*, 2019] and TIPIM [Yang *et al.*, 2021], and two GAN-based attack methods: AdvMakeup [Yin *et al.*, 2021] and AMT-GAN [Hu *et al.*, 2022] as attackers, and ArcFace [Deng *et al.*, 2019] as victim FRS to produce adv-faces on gradient-based attacks. All the gradient-based attacks are applied on LFW and CelebA-HQ datasets to generate adv-faces, while AdvMakeup and AMT-GAN are on the CelebA-HQ dataset.

Detector	FGSM	PGD	DIM	TIM	AdvM.
FGSM	100	99.9	99.7	49.2	0.0
PGD	99.4	100	99.7	6.8	0.0
DIM	98.7	99.4	99.6	2.2	0.0
TIM	71.3	85.8	79.5	93.2	0.0
AdvM.	22.1	19.9	19.5	23.2	98.5

Table 2: Detection accuracy (%) on adv-faces generated by various of attack algorithms. All adv-faces are generated at $\epsilon = 5/255$. Networks are trained on adv-faces generated by attacks in the left column, and tested on adv-faces produced by attacks in the top row.

Implementation details We modify an ImageNet-pretrained XceptionNet [Deng *et al.*, 2009; Chollet, 2017] as the backbone network in our method. We set $N = 7$ to produce a 7×7 feature map in the last convolution layer and choose ReLU as an activation function. We utilize DLIB [Sagonas *et al.*, 2016] for face extraction and alignment, Torchattacks [Kim, 2020] for generating adv-faces, and OpenOOD [Girish *et al.*, 2021] for network training. All face images are aligned and resized to 256×256 before training and testing. The perturbation magnitude ϵ in self-perturbations and adv-faces producing is set to $5/255$, a small value. Threshold γ in the convex hull of gradient image in Algorithm 2 is set to 50. The regularization loss weight β is set to 0.1. Training epochs are set to 5 and convergence is witnessed.

Baselines Previous adversarial example detection methods [Ma *et al.*, 2018; Tian *et al.*, 2021] focus on specific tasks or gradient-based attacks and hence can hardly be effectively extended to GAN-based adv-faces, while face forgery detection methods [Luo *et al.*, 2021; He *et al.*, 2021] are used to detect GAN-made fake faces. On the other hand, some OOD detection methods [Liang *et al.*, 2017; Lee *et al.*, 2018; Sun *et al.*, 2021] only rely on output features and logits of the backbone network and keep agnosticism to unknown attacks, closing to our setting. On account of this, we compare our method to various methods in three problem settings. All baselines listed in Table 1 include methods of adversarial example detection, OOD detection, and face forgery detection. It is worth noting that only our method does not require either

pre-computed adv-faces or access to FRS.

4.1 Assumption Validation

Our assumption for gradient-based noise patterns stands on the visual similarities between adversarial noises. However, the human eye is sometimes unreliable because we cannot observe tiny differences between pixels. To verify the reliability of the hypotheses, we train a simple XceptionNet with real faces from the CelebA-HQ dataset and adv-faces generated by one attack algorithm and test detection accuracy on 1,000 adv-faces per attack. As shown in Table 2, a detector for a specific gradient-based attack is able to generalize to other attacks except those based on GANs. but the success rate on some attacks such as TIM is much lower. We also calculate the Fréchet Inception Distance of various attacks. What the results demonstrate is that attack noises are to some extent similar to each other, and the way of extracting a universal noise pattern is feasible. The result also indicates that noise patterns of GAN-based attacks are not close to gradient-based attack noise patterns and need to be specifically treated.

4.2 Main Results

Gradient-based adv-face detection We first compare the detection performance of our method to other detectors on gradient-based attacks. Baselines include 2 detectors proposed for adversarial example detection [Ma *et al.*, 2018; Tian *et al.*, 2021] and 3 detectors for OOD detection [Liang *et al.*, 2017; Lee *et al.*, 2018; Sun *et al.*, 2021]. The detectors are trained on real images, along with adv-faces images (LID and SID) or pseudo adv-faces images (Ours). We compute the classification AUC for all methods on a dataset comprising of 1k real images and 1k adversarial face images per attack type in LFW and CelebA-HQ datasets. As shown in Table 3, trained on the known attack (FGSM), previous adversarial example detection methods is difficult to detect unknown attacks. In a contrast, our method reaches a high level of detecting gradient-based adv-faces and almost approaches saturation performance on the LFW dataset. This is likely because the network learns a generic representation for adv-faces. The result that the AUC score on LFW is higher than which on CelebA-HQ may contribute to the fact that CelebA-HQ is a more complicated dataset with high resolution and diversified background so self-perturbation works much harder on it. Thus we speculate that a more complex and more diversified data environment will increase the difficulty of adv-face detection tasks.

GAN-based adv-face detection To investigate the advantage of our method on detecting GAN-based samples, we make comparisons with SOTA methods of face forgery detection [He *et al.*, 2021; Luo *et al.*, 2021] and OOD detection [Liang *et al.*, 2017; Lee *et al.*, 2018; Sun *et al.*, 2021]. We train our model with only real faces and self-perturbed faces from CelebA-HQ and test both our approaches and baseline models on GAN-based adv-faces. As shown in Table 4, the performance of our method exceeds which of other algorithms and models. This result verifies our assumption that GAN-made adv-faces have manipulated clues like an abnormal color aberration in high-frequency regions. The observation that detection performance on Adv-Makeup is higher than which on AMT-GAN is likely because abnormal local aberrations of Adv-Makeup are more obvious than that of AMT-GAN. Besides, the result shows that the detection of face forgery can not generalize to adv-faces although both deepfake faces and adv-faces are generated by GANs. This is possible because a huge difference exists between the fingerprints of deepfake GANs and that of adv-face GANs.

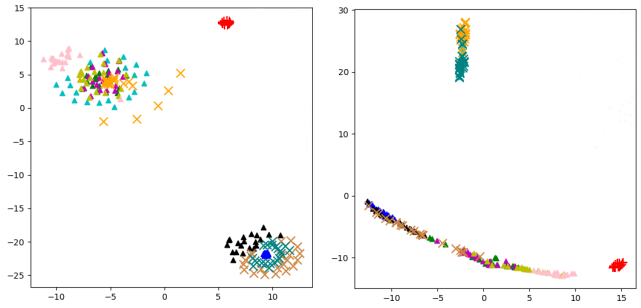

Figure 3: 2D t-SNE visualization. Left: flattened noise images of adv-faces and self-perturbed faces. Right: face features prior to the logit layer extracted by our model, where ▲ denotes adv-face, × denotes self-perturbed face, and + denotes real face. Different colors represent different types of attack and self-perturbation. We use a zero-vector to represent the noise of real faces.

4.3 Ablation

Component ablation As argued, self-perturbation makes the detector learn a more generic representation of adv-faces, Max-Pooling Classifier captures abnormal local color aberrations, and the regularization term helps to boost adv-faces detection. We then conduct an ablation study to verify the effectiveness of each component. As result shown in Table 5, detectors trained without self-perturbation totally failed in detecting adv-faces from TIPIM and GAN-based attacks, indicating the indispensability of self-perturbation in detecting advanced adversarial attacks. There is also an obvious gap between using and not using Max-Pooling Classifier, especially on GAN-based adv-faces. Narrowing the decision boundary, the regularization term helps to filter out a small number of adversarial samples which are not similar enough to self-perturbed faces, further improving detection performance. As for some gradient-based adv-faces such as DIM, detection accuracy is high enough regardless of using Max-Pooling Classifier and a regularization term. It is likely because self-perturbation is similar enough to attack noise patterns so that the representations learned by networks are generic enough.

Complementarity of self-perturbations To explore the necessity and complementarity of self-perturbations, we train a simple XceptionNet on self-perturbed faces generated from CelebA-HQ using only a single self-perturbation. The results are shown in Table 6. As we can see, training a detector only relying on one of the self-perturbations is insufficient for detecting all unseen adv-faces. Due to the differences in generation procedures, self-perturbations are complementary to each other in the detection tasks.

Method	LFW								CelebA-HQ							
	FGSM	BIM	PGD	RF.	MIM	DIM	TIM	TIP.	FGSM	BIM	PGD	RF.	MIM	DIM	TIM	TIP.
LID	76.7	74.0	70.7	73.0	77.7	70.2	62.1	69.0	82.0	55.2	52.5	54.4	57.7	53.7	54.0	59.3
SID	99.7	81.8	73.7	77.8	90.1	72.2	73.4	88.5	96.7	79.2	63.4	72.8	84.5	76.4	81.0	85.2
ODIN	75.6	71.1	71.6	75.2	79.8	73.6	74.8	71.4	76.7	75.7	75.8	75.7	75.0	76.0	72.4	72.2
MD	95.2	91.3	91.9	91.4	90.1	94.0	88.9	86.6	94.1	91.6	91.6	91.7	90.7	93.2	91.9	87.4
ReAct	92.3	89.2	88.4	89.1	89.9	92.2	91.3	90.8	93.6	90.7	90.5	90.6	89.9	92.3	91.9	87.3
Ours	**100**	**100**	**100**	**100**	**99.9**	**99.7**	**100**	**100**	**99.7**	**99.0**	**99.6**	**99.4**	**98.1**	**99.5**	**96.2**	**91.7**

Table 3: Comparison of AUC scores (%) of detecting gradient-based adv-faces from LFW and CelebA-HQ datasets. Detectors of LID and SID are trained on FGSM adv-faces. As for other detectors, all attacks are unseen. "RF." denotes RFGSM and "TIP." denotes TIPIM.

Method	Adv-Makeup	AMT-GAN	Mean
He et al.	52.5	88.2	70.4
Luo et al.	61.8	65.1	62.0
ODIN	63.3	69.7	66.5
MD	72.3	78.2	72.3
React	77.6	82.9	80.3
Ours	**96.6**	**89.7**	**93.2**

Table 4: Comparison of AUC (%) of detecting GAN-based adv-faces from CelebA-HQ dataset. All attacks are unseen by all detectors. The detector of He et al. is pre-trained on CelebA-HQ and detector of Luo et al. is pre-trained on FF++ [Rössler *et al.*, 2019].

Ablation	FGSM	DIM	TIM	TIP.	AdvM.	AMT.
w/o SP	100	92.5	72.4	66.1	52.1	50.8
w/o MPC	99.2	99.0	**96.8**	88.5	90.0	82.0
w/o LU	**99.7**	99.0	95.5	91.2	95.1	88.4
SP+MPC+LU	**99.7**	**99.5**	96.2	**91.7**	**96.6**	**89.7**

Table 5: AUC (%) comparison on CelebA-HQ dataset for ablation study. "w/o SP" means training on adv-faces generated by FGSM instead of self-perturbation. "MPC" refers to Max-Pooling Classifier. "LU" refers to training under the regularization $\mathcal{L}_{\text{uncertainty}}$. "AdvM." and "AMT." denote AdvMakeup and AMT-GAN respectively.

4.4 Analysis of Our Approach

Functionality of self-perturbations To explore the functionality of self-perturbations for gradient-based attacks, we extract noise features and face features of adversarial, self-perturbed, and real faces and visualize them using 2D t-SNE projection. As Figure 3 shows, self-perturbation is very close to attack noise and far from zero-vector (real faces). By separating real faces from self-perturbed faces, the trained network is able to distinguish between real and adv-faces indirectly. Although some adv-faces generated by TIM are not covered by self-perturbations, a model trained on decision boundary regularization still separates them from real faces.

Impact of hyper-parameter Another experiment is about the choice of ϵ. We train a simple XceptionNet on FGSM adv-faces at $\epsilon = 1/255$ to $\epsilon = 15/255$ and test on adv-faces at different ϵ. The result shows that detectors trained on a smaller ϵ are able to detect adv-faces generated with a larger ϵ. But an extremely small ϵ may cause failure in model training. In practice, we do not need to select an extremely small value for ϵ because the attack success rate is too low to take into account as shown in the Appendix. In consideration of this, we choose $\epsilon = 5/255$ (attack success rate lower than 1%) for producing and detecting adv-faces.

Self-Perturbation	FGSM	TIM	TIP.	AdvM.	AMT.
Point-wise	99.6	64.5	68.6	50.8	50.1
Block-wise	81.4	98.9	81.7	76.5	62.3
GC	65.5	70.7	68.0	85.5	78.0

Table 6: Detection AUC (%) on CelebA-HQ. Self-perturbed faces are generated using one of the self-perturbations. "GC" denotes self-perturbation for GAN-based attack mentioned in Algorithm 2.

5 Conclusion

In this paper, we investigate the intrinsic similarities of recent adv-faces and heuristically design three kinds of real face self-perturbations close to attack noise pattern. Regarding adv-faces as OOD data, we propose an FRS-agnostic and attack-agnostic cascaded system for adv-faces detection, which includes real face self-perturbations, decision boundary regularization, and Max-Pooling Classifier focusing on abnormal local color aberrations. Trained on only real faces and self-perturbed real faces, our model learned generic representations for adv-faces. Comprehensive analysis validates the proposed assumptions that noises of adv-faces have intrinsic similarities and exist in high-frequency areas, and extensive experiments demonstrate the improved effectiveness of our method compared with several recent baselines. We also observe that the latent representation learned by the detector is hard to generalize to other domains. For future work, we will try to solve this matter.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61972169, in part by the National key research and development program of China(2019QY(Y)0202, 2022YFB2601802), in part by the Major Scientific and Technological Project of Hubei Province (2022BAA046, 2022BAA042), in part by the Research Programme on Applied Fundamentals and Frontier Technologies of Wuhan(2020010601012182) and the Knowledge Innovation Program of Wuhan-Basic Research, in part by China Postdoctoral Science Foundation 2022M711251.

References

[Avarikioti *et al.*, 2019] Georgia Avarikioti, Kenan Besic, Yuyi Wang, and Roger Wattenhofer. Online payment network design, 2019.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Deng *et al.*, 2019] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[Deng *et al.*, 2021] Zhijie Deng, Xiao Yang, Shizhen Xu, Hang Su, and Jun Zhu. Libre: A practical bayesian approach to adversarial detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 972–982, 2021.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[Dong *et al.*, 2019] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4307–4316, 2019.

[Du *et al.*, 2022] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

[Duda and Hart, 1973] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Willey & Sons, New Yotk, 1973.

[Dumbre *et al.*, 2016] Sudarshan Dumbre, Shamita Kulkarni, Devashree Deshpande, and P.V.Mulmule. Face detection and recognition for bank transaction. *Journal of emerging technologies and innovative research*, 3:108–112, 2016.

[Feinman *et al.*, 2017] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[Gary *et al.*, 2007] B. Huang Gary, Ramesh Manu, Berg Tamara, and Learned-Miller Erik. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, 2007.

[Girish *et al.*, 2021] Sharath Girish, Saksham Suri, Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14074–14083, 2021.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[He *et al.*, 2021] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2534–2541. International Joint Conferences on Artificial Intelligence Organization, 2021. Main Track.

[Hu *et al.*, 2022] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15014–15023, 2022.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[Kim, 2020] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[Kurakin *et al.*, 2016] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[Lee *et al.*, 2018] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[Li *et al.*, 2020] Lixiang Li, Xiaohui Mu, Siying Li, and Haipeng Peng. A review of face recognition technology. *IEEE Access*, 8:139110–139120, 2020.

[Liang *et al.*, 2017] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[Luo *et al.*, 2021] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16321, 2021.

[Ma *et al.*, 2018] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality, 2018.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Ou *et al.*, 2017] Xinyu Ou, Hefei Ling, Han Yu, Ping Li, Fuhao Zou, and Si Liu. Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy. *ACM Trans. Intell. Syst. Technol.*, 8(5), 2017.

[Rössler *et al.*, 2019] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[Sagonas *et al.*, 2016] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

[Shi *et al.*, 2020] Yuxuan Shi, Hefei Ling, Lei Wu, Jialie Shen, and Ping Li. Learning refined attribute-aligned network with attribute selection for person re-identification. *Neurocomputing*, 402:124–133, 2020.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[Sun *et al.*, 2021] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

[Tian *et al.*, 2021] Jinyu Tian, Jiantao Zhou, Yuanman Li, and Jia Duan. Detecting adversarial examples from sensitivity inconsistency of spatial-transform domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9877–9885, 2021.

[Tramèr *et al.*, 2018] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[Wu *et al.*, 2021] Weibin Wu, Yuxin Su, Michael R. Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9020–9029, 2021.

[Yang *et al.*, 2021] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3877–3887, 2021.

[Yin *et al.*, 2021] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1252–1258. International Joint Conferences on Artificial Intelligence Organization, 2021.

[Yu *et al.*, 2019] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7555–7565, 2019.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Zhang *et al.*, 2021] Baiyan Zhang, Hefei Ling, Jialie Shen, Qian Wang, Jie Lei, Yuxuan Shi, Lei Wu, and Ping Li. Mixture distribution graph network for few shot learning. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1, 2021.